An Improved YOLOv5-Based Method for UAV Object Detection

Yunfei Han[a,b], *Abudoula Aierken*[a,b], *Yi Wang*[a,b]* *and Yupeng Ma*[a,b]
hanyf@ms.xjb.ac.cn, abdl@ms.xjb.ac.cn, wangyi@ms.xjb.ac.cn and ypma@ms.xjb.ac.cn

[a] The Xinjiang Technical Institute of Physics & Chemistry
Chinese Academy of Sciences
Urumqi, China

[b] Xinjiang Laboratory of Minority Speech and Language Information Processing
Chinese Academy of Sciences
Urumqi, China

***Abstract*—Object detection based on unmanned aerial vehicle (UAV) images is very challenging. The multi-scale size and high density of objects in the UAV view bring great difficulties. To fully address this issue to unleash the potential of UAV applications, the YOLOv5-STD model is proposed. First, add one more head to locate extremely small object detection by shallow image features; second, use the attention mechanism to optimize the backbone by the transformer; third, use SPD-Conv to avoid the loss of fine-grained image feature information. At the last, sufficient experiments on the dataset VisDrone 2022 have proven that the model has good performance, compared with the basic model, the improved model has an average improvement of about 7% in mAP@.5 metrics, and the ablation experiments have verified that its improvement skills have a positive effect on the model. This paper can help developers and researchers get a better experience in the analysis and processing of unmanned aerial vehicle images.**

Keywords-object detection; yolov5; transformer; space-to-depth convolution

I. INTRODUCTION

Intelligent applications based on unmanned aerial vehicle image processing technology have been widely used in various industries, such as intelligent monitoring, such as intelligent surveillance[1], search and rescue[2], infrastructural inspection[3], geographical mapping[4], agricultural hazard prevention[5], etc. These applications all need to fully understand the scene environment in the image. The first problem to be solved is object detection, including recognizing what categories of objects are in the scene, and locating where these objects are. Due to the large change of UAV height, the particularity of shooting angle and position, and the large number of positive and negative samples covered in the image, object detection in UAV images is faced with great difficulties such as small objects, different shooting angles, complex backgrounds, and high object density, forming a unique small object detection challenge, and further research is needed to improve the accuracy of object detection and the understanding ability of UAV images.

In recent years, with the rise of deep learning and the availability of large-scale labeled samples(e.g., Pascal VOC[6] and COCO[7]), many state-of-the-art object detectors based on deep learning have been proposed in the past ten years, and have been very successful in the field of computer vision. Since the beginning when Con-volution Neural Networks (CNNs)[8] were successfully introduced in object detection, Sermanet et al. presents a one-stage framework OverFeat[9] object detection based on CNNs. Up to now, excellent object detectors based on deep learning have been proposed, including R-CNN[10], Fast R-CNN[11], YOLO[12-15], Faster R-CNN[16], SSD[17], R-FCN[18], RetinaNet[19], CornerNet[20], CenterNet[21, 22], etc.

Although the above models have achieved good results on public datasets, they are inadequate for the detection of small objects in UAV images. Many scholars have conducted research in different dimensions in response to the challenge of small object detection. For example, based on multi-scale or multi-level feature fusion to enhance feature acquisition of small objects [22-28]; based on context information and attention mechanisms to enhance the perception of small objects [29-32]; based on super-resolution techniques to enhance the resolution of small objects and transform them into medium or large objects for detection [33-35]; based on cascaded multiple detectors [36, 37] or methods based on image or feature patch[38, 39] for multiple detection fusion.

In this paper, **YOLOv5-STD**(YOLOv5 with **S**mall Object Detection Head, **T**ransformer, and Space-to-**D**epth Convolution module) based on YOLOv5 is proposed, and its focus is on the integration of multiple skills which may strengthen the perception of small object features for improving accuracy. Specifically, we add one more head for small object detection, then we use a transformer [40] to optimize the backbone, and followed by we replace the original convolution with space-to-depth Convolution (SPD-Conv) [41] to explore the prediction potential. Finally, the experimental results show that our method achieves significant improvement for small object detection on the VisDrone dataset, which is also competitive compared with the state-of-the-art methods. The overall framework of YOLOv5-STD is demonstrated in Figure. 1, which will be introduced in detail in Section 3.

The contributions are listed as follows:

* Corresponding author: Yi Wang(wangyi@ms.xjb.ac.cn)

1. We add one more head for small object detection, which can localize small objects by shallow image features.

2. We use the attention mechanism to optimize the backbone by the transformer.

3. We use SPD-Conv to avoid the loss of fine-grained image feature information.

4. On the VisDrone 2022 dataset, our proposed YOLOv5-STD achieves 41.90% mAP. Experiments have shown that the fusion of multiple small object detection tricks can be very effective.

II. Related Work

A. Object Detectors

Most of these excellent object detectors have made breakthrough progress and have achieved state-of-the-art results on public datasets. As the first representative of two-stage object detection based on deep learning, R-CNN[10] first obtained the features of candidate regions by CNN automatically. Faster R-CNN[16] formed the basic framework of object detection, and many algorithms are extended based on it. Its appearance opened a new chapter of object detection. One of its most outstanding contributions is to use CNNs to generate region proposals with anchor boxes in the whole infer process. Two-stage detectors with lots of region proposals require huge computation and run-time memory. In contrast, YOLO series models[12-15], SSD[17], and RetinaNet[19] as one-stage detectors alleviate the problem of inference efficiency effectively. One-stage detectors directly treat object detection as regression problems by taking input images and learning categories probabilities and bounding box coordinates. One-stage detectors are more likely to infer faster than two-stage detectors. CenterNet opened a new idea of anchor-free object detection. Not only the object categories probabilities were predicted by the image features, but also the coordinates of the bounding box key points were predicted directly. These met eliminated the dependence on anchor boxes and promoted the development of anchor-free object detection.

Up to now, YOLO series models have evolved and developed with excellent performance. They are the typical representative masterpieces in the field of object detection. In this work, the baseline model is YOLOv5[15], it consists of three parts: backbone, neck, and head. The backbone is based on convolutional neural networks to fully extract the depth features of images; the neck is designed to make better use of the features extracted by the backbone at different levels; the head is used to predict the class and bounding box for the object. YOLOv5 has several advantages. It uses mosaic data augmentation, which enriches the small object samples in the dataset and makes the detection network more robust. In the backbone, YOLOv5 builds C3 layers by improving the CSPBottleneck, which is simpler, faster, lighter, and achieves better results at a nearly similar loss. In the neck, YOLOv5 uses the SPP module to fuse multi-scale features, improve the perceptual field, and enrich the expressiveness of the feature map, which is beneficial in the case of large differences in object sizes in images. The YOLOv5 project has a very clear architecture, and rich engineering support functions and its easy and efficient deployment makes it the best choice for engineering projects. In summary, we choose YOLOv5 as the baseline model to improve and optimize UAV object detection.

B. Object Detectors Methods for Small Object Detection

Recently, a lot of work has been done in small object detection optimization research. Multi-level or multi-scale features are used to enhance the fine-grained feature representation of small objects. EfficientDet[26] proposed a weighted bi-directional pyramid network (BiFPN), which adds efficient bi-directional cross-scale link and weighted feature fusion to the FPN network, thus enabling convenient and fast multi-scale feature fusion. Context and attention are used to enhance the perception of small objects. FA-SSD[31] uses feature fusion to obtain contextual information about small objects to extract shallow features from small objects that lack semantic information and uses an attention module to allow the network to focus only on important parts. The super-resolution technique makes it possible to transform small objects into bigger ones. JCS-Net[33] studies the relationship between large-scale and small-scale pedestrians based on the super-resolution network, which is responsible for amplifying the small object by upsampling and recovering the details of the small-scale pedestrians to obtain an amplified object. Combining the super-resolution loss with the classification loss, the reconstructed small-scale object contains both the original and output information of the super-resolution network. Cascade R-CNN[36] uses cascade regression as a resampling mechanism to increase the IoU value of proposals stage by stage so that the resampled proposals from the previous stage can be adapted to the next stage with a higher threshold. In MPFP-Net[39], features are sliced into patches, and these patches are divided into class-affiliated subsets, to which the patches are related. The network contains bottom-up and crosswise connections to fuse the features of different scales to achieve better accuracy.

III. YOLOv5-STD

A. Overview

The basic framework of YOLOv5 can be divided into 4 parts: Input, Backbone, Neck, and Prediction.

The Input part enriches the dataset by stitching data augmentation, which requires low hardware equipment and low computational cost. However, it will cause the original small objects in the dataset to become smaller, resulting in a decrease in the generalization performance of the model.

The Backbone part mainly consists of CSP modules, and feature extraction is performed by CSPDarknet53.

FPN and Path Aggregation Network (PANet) are used in Neck to aggregate the image features in this stage.

Finally, the network performs object prediction and passes the predicted output.

YOLOv5 has five different versions including YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x. The main difference between the different versions is the model depth and width. The original YOLOv5 was modified to specialize in small object detection. YOLOv5n is the smallest version of

the YOLO series, more suitable for deployment on a variety of hardware platforms, and its architecture is simpler and clearer. Figure. 1 demonstrates the framework of the YOLOv5-STD.

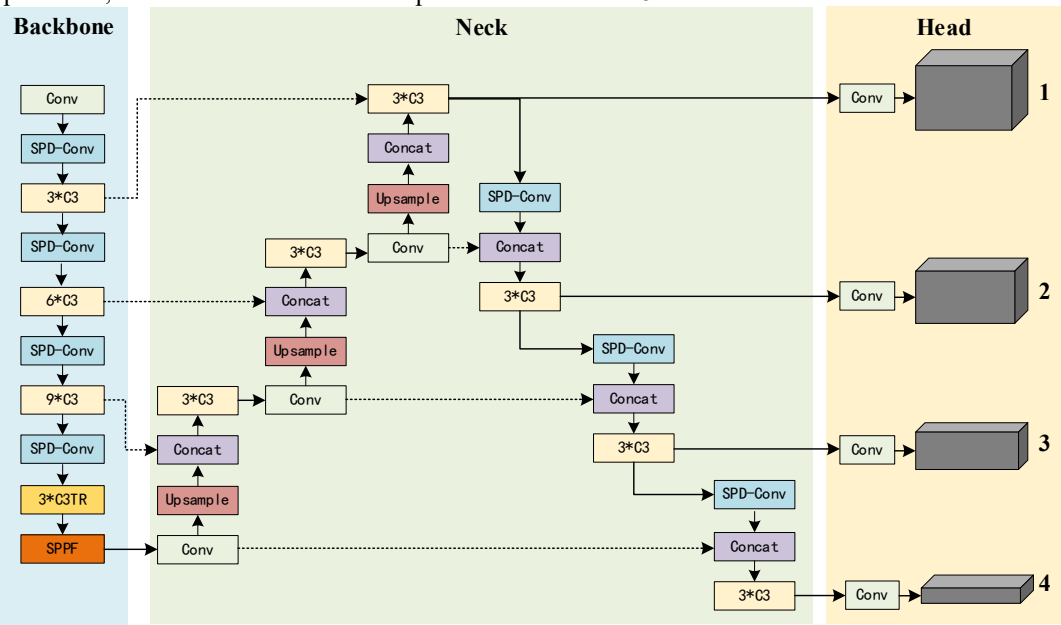

Figure 1. The structure of YOLOv5-STD.

B. *Small Object Detection Head*

We found a large number of small objects in the VisDrone 2022 dataset. As shown in Figure 2, the proportion of small objects smaller than 32*32 and larger than 8*8 is the largest, accounting for 55.69%. At the same time, there is also a certain proportion of extremely small objects smaller than 8x8, accounting for 6.65%. In response to the problem of a large number of small objects in drone images, adding a prediction head for small object detection can better cope with multi-scale object detection in drone scenes overall. As shown in Figure 1, the new prediction head (Head 1) utilizes high-resolution image features and combines them with lower-level visual feature maps to perceive small objects more efficiently. By adding a prediction head for extremely small objects, the detection performance can be greatly improved, although the computational complexity and resource consumption of the model has increased.

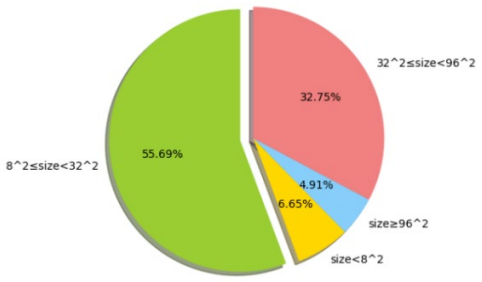

Figure 2. The proportion of different sizes of objects.

C. *Transformer*

The transformer model is a revolutionary model proposed by Google, not only in machine translation but also in text summarization, speech recognition, question-answering systems, dialogue systems, and machine vision. It uses a completely attention-based approach that makes training and inference much faster, while also improving performance.

The core idea of the transformer model is the self-attention mechanism, which can automatically learn and focus on important information in the input sequence, achieving interaction of information at different positions in the sequence. The self-attention mechanism in the Transformer model adopts multi-head attention, which can simultaneously focus on different subspaces of the input sequence, thereby enhancing the model's expressive power. Compared with traditional RNN models, the Transformer model has the following advantages:(1) Avoids time-series calculations in RNN models, can process input sequences in parallel, and makes training and inference much faster; (2) Achieves interaction of information at different positions in the sequence through self-attention mechanisms, enhancing the model's expressive power and allowing it to handle longer input sequences; (3) The Transformer model uses techniques such as residual connections and layer normalization, making the model more stable and easier to train.

Inspired by the vision transformer [40], we replaced the Bottlenecks in the last C3 blocks in the original version of YOLOv5 with a transformer block. It can not only extract local features but also use attention mechanisms to pay attention to the region where small objects are located. It can also explore the feature representation potential with the self-

attention mechanism. The transformer encoder blocks have better performance on occluded objects with high density.

D. Space-to-depth Convolution

The space-to-depth convolution is a special type of convolution operation, which is widely used in the field of deep learning. It is mainly used for image processing tasks, such as image classification, object detection, posture estimation, motion recognition, and so on. It converts the input image data from spatial dimensions to depth dimensions, thus increasing the nonlinearity and sparsity of the network, and improving the representation ability and computational efficiency of the model. Specifically, space-to-depth convolution divides the input image data into multiple blocks, arranges the pixels in each block in the depth dimension, and then combines these blocks in a certain order into new image data. This operation can reduce the spatial dimension of input data while increasing the depth dimension, thus enabling the network to better detect objects of different sizes, improving the efficiency and accuracy of object detection. In this paper, we replace most of the convolution in Yolov5 with space-to-depth convolution to better recognize multiscale objects and more accurate object detection in unmanned aerial vehicle images.

IV. Experiments

A. Datasets and Experiment settings

The VisDrone2022-DET dataset, which is the same as the VisDrone2019-DET dataset and the VisDrone2018-DET dataset, consists of 7,019 static photos taken by drone platforms in various locations and at various heights [18]. The test-dev set has 1610 images, and the train and val set each has 6,471 and 548 images. Images are labeled and annotated with bounding boxes and ten predefined classes (i.e., pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle). All models in this study are tested on the test-dev set after being validated on the val set and trained on the train set. Finally, we show the performance of object detection on the test-dev set and compare it with the base object detection models.

B. Implementation Details

We implement YOLOv5-STD on Pytorch 1.7.1. All of our models use an NVIDIA K80 GPU for training and testing. In the training phase, we use part of the pre-trained model of yolov5 (yolov5n, yolov5s, yolov5m, yolov5l, yolov5x), because YOLOv5-STD and YOLOv5 share part of the backbone and some part of the head. We use SGD optimizer for training, and the training hyperparameters were set to an initial learning rate of 1e-2, a momentum of 0.98, weight decay of 0.001, warm-up epochs of 5, and warm-up momentum of 0.95, and the NMS (Non-Maximum Suppression) threshold was also set as 0.6 in all experiments. The size of the input image of our model is 640*640 pixels. We set the batch size to 32, 16, 16, 8, 4 for Yolov5n-STD, yolov5s-STD, yolov5m-STD, yolov5l-STD, yolov5x-STD. All models were trained on VisDrone2022 train set for 500 epochs with early stopping patience of 100 epochs.

C. Comparison with Base Models

Due to the submission limit of the VisDrone2022 competition server, we only obtained results for the five base models and the five improved models on the testset-challenge. The test results are shown in TABLE I. The larger the model, the richer the image features that can be obtained and the better its object detection results. The improved models are about 5 percentage points higher than the base models in terms of mAP@.5 and mAP@.5:.95 indicators, indicating that the improvement based on STD has better effectiveness and stability.

TABLE I. Base models and improved models test results

Methods	**mAP @.5(%)**	**mAP @.5:.95(%)**
Yolov5n	23.0	11.6
Yolov5s	*28.7*	15.5
Yolov5m	*32.1*	18.1
Yolov5l	34.4	19.9
Yolov5x	35.2	20.5
Yolov5n-STD	31.6	17.2
Yolov5s-STD	36.1	20.2
Yolov5m-STD	39.0	22.5
Yolov5l-STD	40.4	23.5
Yolov5x-STD	41.9	24.5

D. Ablation Study

To fully verify the effectiveness of the improvement based on the small object detection head, transformer, and space-to-depth convolution module, ablation experiments were conducted for the three major modules. Based on the stability of STD, to save model calculations and improve experimental efficiency, the contribution of each module to the model is verified based on the minimized Yolov5n model. By recombining the three modules in the model and testing on the testset-challenge, the results obtained are shown in Table II, and the comparison of detection results between Yolov5n and Yolov5n-STD models is shown in Figure 3.

TABLE II. Test results of improved models based on different combination methods

No.	Methods	mAP @.5(%)	mAP @.5:.95(%)
1	Yolov5n	23.0	11.6
2	Yolov5n-small	27.8	14.6
3	Yolov5n-spd	25.0	12.8
4	Yolov5n-tr	23.7	11.9
5	Yolov5n-small-spd	30.7	16.6
6	Yolov5n-small-tr	28.7	15.3
7	Yolov5n-spd-tr	25.9	13.4
8	Yolov5n-STD	31.6	17.2

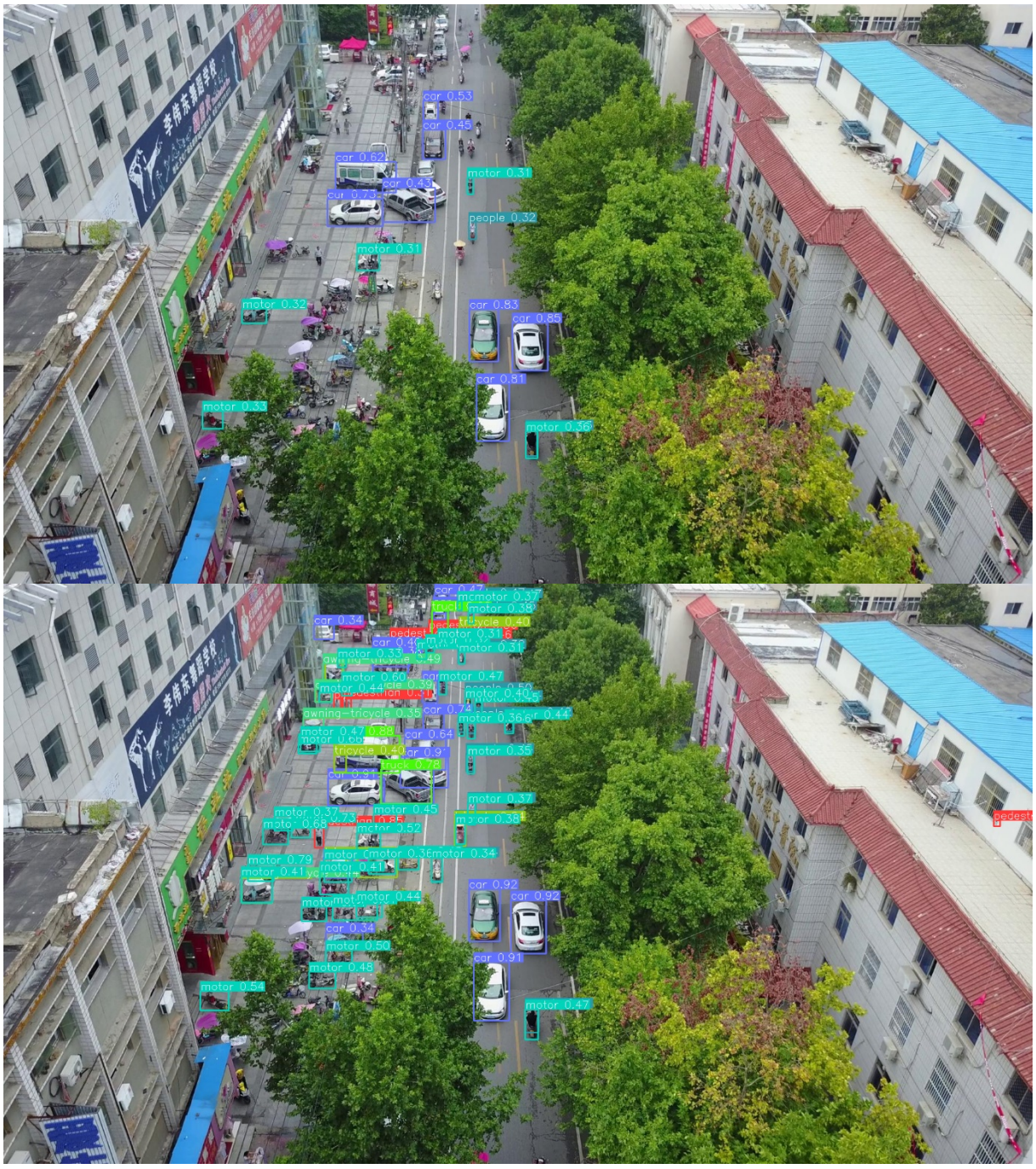

Figure 3. Comparison of test results between Yolov5n and Yolov5n-STD models.

1) ***Effect of Small Object Detection Head.*** The new small object detection head utilizes high-resolution image features and combines them with lower-level visual feature maps to perceive small objects more efficiently. We added a small object detection head to Yolov5n, Yolov5n-spd, Yolov5n-tr, and Yolov5n-spd-tr, from the results in TABLE III, improved models mAP@.5 increase by about 5% on average, mAP@.5:. 95 increase by about 4%. The small object detection head has the greatest positive impact on model performance, and the computational complexity and resource consumption it brings is very worthwhile.

2) ***Effect of Transformer.*** The transformer blocks can not only extract local features but also use attention mechanisms to pay attention to the region where small objects are located. Due to the GPU memory limitations, we only add one transformer block to the last C3 block. We also improved four models: Yolov5n, Yolov5n-spd, Yolov5n-small, and Yolov5n-small-spd, the results are shown in

TABLE IV, improved models mAP@.5 increase by about 1% on average, mAP@.5:. 95 increase by about 0.5%. Only one transformer module has been added to the models, which still has a slightly positive impact. This can prove that the transformer block is useful in small object detection, and its significant results need to be further verified on a larger memory GPU.

*3) **Effect of Space-to-Depth Convolution.*** Space-to-depth convolution firstly increases the nonlinear representation ability of the network, which can better identify multi-scale objects. Secondly, it increases the sparsity of the network and improves computational efficiency. It replaced the most of convolutions in four models: Yolov5n, Yolov5n-spd, Yolov5n-small, and Yolov5n-small-spd with space-to-depth convolutions, the results are shown in TABLE V, improved models mAP@.5 increase by about 2.5% on average, mAP@.5:. 95 increase by about 1.5%. The space-to-depth convolution effectively improves model performance while optimizing the computational complexity, and is very effective in small object detection.

TABLE III. ABLATION STUDY ON SMALL OBJECT DETECTION HEAD

No.	Base Method	With Small Object Detection Head	mAP @.5(%)	mAP @.5:.95(%)
1	Yolov5n	Yolov5n-small	+4.8	+3.0
2	Yolov5n-spd	Yolov5n-small-spd	+4.7	+3.8
3	Yolov5n-tr	Yolov5n-small-tr	+5.0	+3.4
4	Yolov5n-spd-tr	Yolov5n-STD	+5.7	+4.8

TABLE IV. ABLATION STUDY ON TRANSFORMER

No.	Base Method	With Transformer	mAP @.5(%)	mAP @.5:.95(%)
1	Yolov5n	Yolov5n-tr	+0.7	+0.3
2	Yolov5n-small	Yolov5n-small-tr	+0.9	+0.7
3	Yolov5n-spd	Yolov5n-spd-tr	+0.9	+0.6
4	Yolov5n-small-spd	Yolov5n-STD	+0.9	+0.6

TABLE V. ABLATION STUDY ON SPD-CONV

No.	Base Method	With Space-to-Depth Convolution	mAP @.5(%)	mAP @.5:.95(%)
1	Yolov5n	Yolov5n-spd	+2.0	+1.2
2	Yolov5n-small	Yolov5n-small-spd	+2.9	+2.0
3	Yolov5n-tr	Yolov5n-spd-tr	+2.2	+1.5
4	Yolov5n-small-tr	Yolov5n-STD	+2.9	+1.9

V. CONCLUSION

This paper proposed an improved small object detector YOLOv5-STD, which is based on YOLOv5. We added some skills to tackle small object detection issues, such as small object detection head, transformer, and space-to-depth convolution. The YOLOv5-STD is especially good at object detection in unmanned aerial vehicle images. On the VisDrone2022-DET dataset, a large number of experiments based on the improved models have shown better detection results, indicating that this method is feasible. In addition, the ablation study experiments fully prove that the small object detection head, transformer, and space-to-depth convolution have a positive effect on small object detection, which proves that the improved methods have good effectiveness and stability. This paper can help developers and researchers get a better experience in the analysis and processing of unmanned aerial vehicle images.

ACKNOWLEDGMENT

This research was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region of China (2020D01B55) and the West Light Foundation of The Chinese Academy of Sciences (2019-XBQNXZ-B-009).

REFERENCES

[1] Y. LI, Han, et al. Multi-block SSD based on small object detection for UAV railway scene surveillance[J]. Chinese Journal of Aeronautics, 2020, v.33;No.171(06):179-187.

[2] L. HONG, Y. WANG, Y. DU. Xin CHEN, Yujun ZHENG. UAV search-and-rescue planning using an adaptive memetic algorithm. Front. Inform. Technol. Electron. Eng, 2021, 22(11): 1477–149.

[3] H. Cheng, J. Yang. Solar Power Plant Maintenance with Thermal UAV Inspection Technol-ogy[J]. Power: The Magazine of Power Generation and Plant Energy Systems, 2022(6):166.

[4] H. Jia , L. Wang , D. Fan. The application of UAV LiDAR and tilt photography in the early identification of geo-hazards[J]. The Chinese Journal of Geological Hazard And Control, 2022, 32(2):60-65.

[5] Lnl A , Jls A , Yc B , et al. Using UAV-based thermal imagery to detect crop water status variability in cotton - ScienceDirect[J]. Smart Agricultural Technology, 2021.

[6] Everingham M , Gool L , Williams C K, et al. The Pascal Visual Object Classes (VOC) Challenge[J]. Springer US, 2010(2).

[7] Lin, T.-Y. et al. Microsoft COCO: Common Objects in Context. in Computer Vision – ECCV 2014 (eds. Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T.) 740–755 (Springer Interna-tional Publishing, 2014).

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Neural Information Processing Systems, pp. 1097–1105, 2012.

[9] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks,"arXiv preprint arXiv:1312.6229, 2013

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14), pp. 580-587, Columbus, Ohio, USA, 2014.

[11] R. Girshick, "Fast R-CNN," in Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV '15), pp. 1440–1448, Santiago, Chile, 2015.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR '16), pp. 779–788, Las Vegas, Nev, USA, 2016.

[13] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, (CVPR '17), pp. 6517–6525, Honolulu, Hawaii, USA, 2017.

[14] J. Redmon, and A. Farhadi, "YOLOv3: an incremental improvement (2018)." arXiv preprint arXiv:1804.02767, 2018.

[15] Glenn Jocher, Alex Stoken, Ayush Chaurasia, et al., ultralytics/yolov5, https://github.com/ultralytics/yolov5, 2022.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,"IEEE Transactions on Pattern Analysis andMachine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.

[17] W. Liu, D. Anguelov, D. Erhan, et al., "SSD: single shot multibox detector," in Proceedings of the Computer Vision – ECCV 2016, vol. 9905 of Lecture Notes in Computer Science, pp. 21–37, 2016.

[18] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks," arXiv preprint arXiv:1605.06409, 2016.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Doll´ar. Focal loss for dense object detection. ICCV, 2017.

[20] H. Law, J. Deng, "CenterNet: Keypoint Triplets for Object Detection", European Conference on Computer Vision, 2018.

[21] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang and Q. Tian, "CenterNet: Keypoint Triplets for Object Detection", arXiv preprint arXiv:1904.08189, 2019.

[22] X. Zhou, D. Wang, P. Krähenbühl, "Objects as Points", arXiv preprint arXiv:1904.07850, 2019.

[23] G. Hu, Z. Yang, L. Hu, et al. Small object detection with multiscale features. International Journal of Digital Multimedia Broadcasting, 2018, 2018.

[24] S. Liu, L. Qi, H. Qin, et al. Path aggregation network for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.

[25] Q. Zhao, T. Sheng, Y. Wang, et al. M2det: A single-shot object detector based on multi-level feature pyramid network. Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 9259-9266.

[26] M. Tan, R. Pang, Q. Le. Efficientdet: Scalable and efficient object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.

[27] X. Yu , Y. Gong , N. Jiang, et al. Scale match for tiny person detection. Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020: 1257-1265.

[28] Y. Gong, X. Yu, Y. Ding, et al. Effective Fusion Factor in FPN for Tiny Object Detection. Proceedings of the IEEE/CVF Winter Conference on Workshop on Applications of Computer Vision. 2021.

[29] L. Guan, Y. Wu, J. Zhao. Scan: Semantic context-aware network for accurate small object detection. International Journal of Computational Intelligence Systems, 2018, 11(1): 951-961.

[30] Y. Yuan, Z. Xiong, Q. Wang. VSSA-NET: Vertical spatial sequence attention network for traffic sign detection. IEEE transactions on image processing, 2019, 28(7): 3423-3434.

[31] J. S. Lim, M. Astrid, H. J. Yoon, et al. Small object detection using context and attention. 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE, 2021: 181-186.

[32] N. Carion, F. Massa, G. Synnaeve, et al. End-to-end object detection with transformers, European conference on computer vision. Springer, Cham, 2020: 213-229.

[33] Y. Pang , J. Cao, J. Wang, et al. JCS-Net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images. IEEE Transactions on Information Forensics and Security, 2019, 14(12): 3322-3331.

[34] Y. Bai, Y. Zhang, M. Ding, et al. SOD-MTGAN: Small Object Detection via Multi-Task Generative Adversarial Network, Proceedings of the European Conference on Computer Vision (ECCV), pp. 206-221, 2018.

[35] Y. Bai, Y. Zhang, M. Ding, et al. Finding Tiny Faces in the Wild With Generative Adversarial Network, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21-30, 2018.

[36] Z. Cai, N. Vasconcelos. Cascade R-CNN: Delving into High Quality Object Detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6154-6162.

[37] Q. Lin, Y. Ding, H. Xu, et al. ECASCADE-RCNN: Enhanced cascade RCNN for multi-scale object detection in UAV images[C]//2021 7th International Conference on Automation, Robotics, and Applications (ICARA). IEEE, 2021: 268-272.

[38] A. Wang, W. Li, X. Wu, et al. MPANet: Multi-Patch Attention For Infrared Small Target object Detection. arXiv preprint arXiv:2206.02120, 2022.

[39] P. Shamsolmoali, J. Chanussot, M. Zareapoor, et al. Multi-patch feature pyramid network for weakly supervised object detection in optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-13.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929, 2020.

[41] R. Sunkara, T. Luo. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. arXiv preprint arXiv:2208.03641, 2022.